

ISSN (print): 2421-5783
ISSN (on line): 2421-5562



Consiglio Nazionale delle Ricerche

IRCFES

ISTITUTO DI RICERCA SULLA CRESCITA ECONOMICA SOSTENIBILE
RESEARCH INSTITUTE ON SUSTAINABLE ECONOMIC GROWTH

Rapporto Tecnico

Numero 6, maggio 2018

Follow the Byterfly
and enjoy open knowledge

GIANCARLO BIRELLO, ANNA PERIN

Direttore Secondo Rolfo


Direzione CNR-IRCRES
Istituto di Ricerca sulla crescita economica sostenibile
Via Real Collegio 30, 10024 Moncalieri (Torino), Italy
Tel. +39 011 6824911 / Fax +39 011 6824966
segreteria@ircres.cnr.it
www.ircres.cnr.it


Sede di Roma Via dei Taurini 19, 00185 Roma, Italy
Tel. +39 06 49937809 / Fax +39 06 49937808

Sede di Milano Via Bassini 15, 20121 Milano, Italy
Tel. +39 02 23699501 / Fax +39 02 23699530

Sede di Genova Università di Genova Via Balbi, 6 - 16126 Genova
Tel. +39 010 2465459 / Fax +39 010 2099826

Redazione Secondo Rolfo (direttore responsabile)
Antonella Emina
Anna Perin
Diego Margon
Isabella Maria Zoppi

 redazione@ircres.cnr.it

 www.ircres.cnr.it/index.php/it/produzione-scientifica/pubblicazioni

RAPPORTO TECNICO CNR-IRCRES, numero 6, maggio 2018



Copyright © maggio 2018 by CNR - IRCRES

Follow the Byterfly and enjoy open knowledge

GIANCARLO BIRELLO^a, ANNA PERIN^b

^a CNR-IRCrES, Consiglio Nazionale delle Ricerche - Istituto di Ricerca sulla Crescita Economica Sostenibile, Ufficio IT, strada delle Cacce, 73 Torino - Italia

^b CNR-IRCrES, Consiglio Nazionale delle Ricerche - Istituto di Ricerca sulla Crescita Economica Sostenibile, Biblioteca, via Real Collegio 30, Moncalieri (TO) - Italia

corresponding author: anna.perin@ircres.cnr.it

ABSTRACT

Can a repository answer to the five laws of Library Science, proposed by S. R. Ranganathan in 1931? This report presents an overview of “Byterfly” repository model, including the adopted metadata, the used software, and the policies.

KEYWORDS: Repository, open access, open source, digital library.

DOI: 10.23760/2421-5562.2018.006

HOW TO CITE THIS ARTICLE

Birello G., & Perin A. (2018) Follow the Byterfly and enjoy open knowledge, *Rapporto Tecnico CNR-IRCrES*, n. 6/2018. <http://dx.doi.org/10.23760/2421-5562.2018.006>

CONTENTS

1	INTRODUCTION.....	3
2	REPOSITORY	3
3	REPOSITORY ARCHITECTURE	4
3.1	Back end server.....	4
3.2	Image Server	4
3.3	Front-end server.....	5
4	THE POLICIES.....	5
4.1	Policies Covered	5
5	COLLECTION HOSTED	6
6	METADATA	6
7	DIGITIZING, RULES AND SUGGESTIONS	9
8	BROWSING AND NAVIGATE	10
9	FINAL CONSIDERATIONS	10
10	REFERENCES.....	11

Follow the Byterfly and enjoy open knowledge

GIANCARLO BIRELLO, ANNA PERIN

1 INTRODUCTION

The “Five laws of library science” is a theory proposed by S. R. Ranganathan in 1931, detailing the principles of operating a library system. Many librarians worldwide accept them as the foundations of their philosophy.

These laws are¹:

- Books are for use.
- Every reader their book.
- Every book its reader.
- Save the time of the reader.
- The library is a growing organism.

Can we apply these laws to a repository?

2 REPOSITORY

A repository is a digital library with a focused collection of digital objects that can include texts, images, audio and video files, stored as electronic media formats combined with metadata that provide information about the resource; repository can immensely vary in size and scope.

IT Office and the library of IRCrES-CNR manage and maintain different repositories since 2012. During 2017, we implemented a brand new repository, “Byterfly”, with the aim of being in line with the first law: “*Books are for use*”. In the modern world, making books available to all readers means uploading the books in an accessible and reachable online site.

Byterfly is an open source repository for both architecture and software. We used Fedora Repository, one of the most popular software for the management of digital assets: it is robust, persistent, and able to manage big data.

For the front-end, we used Dupal CMS and Islandora, that is a repository manager and repository model based, with workflow and modules expandable and customizable for a lot of media type. For the search functions, we used Solr.

The metadata (data of data) adopted are simple Dublin Core. The repository is OAI-PMH compliant, as a primary key for visibility and dissemination through meta-repositories, but also simply by web crawler search engine.

¹ From Wikipedia (https://en.wikipedia.org/wiki/Five_laws_of_library_science)

3 REPOSITORY ARCHITECTURE

The architecture is based on a virtualized platform implementing KVM open-source hypervisor² running on top of Ubuntu server 16.04 in conjunction with OpenVSwitch to manage networking and iSCSI with multipath to manage SAN storage.

Current architecture requires three virtual machines: a) the repository to manage digital objects, b) the image server to render image files and c) the front-end, hosting the interface for public access to the repository objects and data. All virtual machines are running Ubuntu server 16.04 LTS.

Finally, a highly reliable storage (SAN) completes the architecture; it supports the repository for objects preservation and the virtual machine for system backups.

3.1 Back end server

The back-end virtual machine hosts Fedora Repository, BlazeGraph and Solr, each one of them running in a separated Java servlet container; the first two, in a couple of Tomcat instances, while Solr is running on self-included Jetty servlet container.

Fedora Repository manages the Byterfly conservation task, which includes API's for programmable ingesting, semantic description of the relationships among the objects and their management based on models.

Unfortunately, the Fedora triple-store manager Mulgara has low performance and poor versatility; that is why we chose to replace it with BlazeGraph, a powerful triple-store manager. You can follow our steps³ to modify Fedora configuration for the new triple-store based on tripi-sail module (thanks to Discovery Garden!).

Finally, Solr is the search and indexing engine, which is in charge of indexing all ingested data stream contents and provides very fast search results. Configuration is customized to index date, full-text data stream and all Dublin Core elements in order to provide right fields to be used with Views and search facet block into Islandora front-end.

3.2 Image Server

The Islandora Large Image and Book modules use Openseadragon and Internet Archive Bookreader to show images, pages and books. These two components are very powerful and provide the front-end with a better user experience for accessing to digital objects. Both of them are based on JPEG2000 image format and they need an image tile source to show properly.

In the past, we used Adore-djatoka servlet to feed the image viewer components with the required derivatives; it run into a container which in most cases was the same container used by Fedora repository. For a few months, Islandora Openseadragon head version has started to support IIIF (International Image Interoperability Framework) standard, while simultaneously Cantaloupe open-source dynamic image server has reached a good maturity and reliability. Moreover, starting on a first implementation of Islandora Internet Archive Bookreader version 2 developed by Diego Pino Navarro⁴, we added it some features and made code ready to be used properly with Cantaloupe IIIF server.

The result has been an Islandora front-end full based on IIIF server, without any need of adore-djatoka. Cantaloupe is a high performance image derivative generator with powerful cache features, so we decided to build a third virtual server in order to host exclusively the IIIF server, which is able to server more than one Islandora at time, i.e. our Cantaloupe server is providing images to three distinct repository front-ends.

Byterfly development code of the repository is available to the open community in dev site⁵.

² <http://dev.digibess.it/doku.php?id=hypervisor:start>

³ http://dev.digibess.it/doku.php?id=reloaded:be_repmulg

⁴ https://github.com/DiegoPino/islandora_internet_archive_bookreader/tree/7.x-2dev

⁵ <http://dev.digibess.it/doku.php?id=reloaded>

3.3 Front-end server

“Open-access” means also an easy and comfortable way to access digital objects and data preserved into the repository. This is one of the key reasons why we chose Islandora many years ago and we confirmed the choice for the Byterfly, too. The Islandora framework is integrated in Drupal CMS running onto a LAMP (Linux, Apache, MySQL, PHP) stack; it is a completely open-source software, as other components, and it is supported by a dynamic and extended community.

Islandora Byterfly implementation differs from previous architectures and standard guidelines for two main aspects: full IIF server support and intensive use of Views. As the previous paragraph highlights, Byterfly replaces Adore-djatoka with Cantaloupe to manage image derivatives for Openseadragon and Internet Archive Bookreader viewers.

The second remarkable issue is the fully replacement of Islandora default collection display by Views, which is also used for some custom navigation bars for a better user’s experience. We developed a large variety of Views able to display collections by year, by decade, by title, making alternatively use of accordion and grouping. Moreover, Views also manages metadata display allowing different layouts depending on object namespace, so that one can customize how Dublin Core are presented to the user.

4 THE POLICIES

OpenDOAR (The Directory of Open Access Repositories) has created a simple tool to help repository administrators to formulate and present their repository’s policies. It provides a series of check boxes and pick lists for all the key policy options, which can be very quickly selected.

The tool provides recommended options for minimum compliance with the aims of the Open Access movement, and for optimizing the use of a repository. For example, the minimum policy recommends allowing re-use of metadata for non-profit purposes, but it prohibits any commercial re-use.

4.1 Policies Covered

The policies cover all aspects of the repository, in particular:

- **Metadata Policy**– for information describing items in the repository.
Access to metadata; Re-use of metadata.
- **Data Policy** – for full-texts and other full data items.
Access to full items; Re-use of full items.
- **Content Policy** – for types of documents and datasets held.
Repository type; Type of materials held; Principal languages.
- **Submission Policy** – concerning depositors, quality and copyright.
Eligible depositors; Deposition rules; Moderation; Content quality control; Publishers’ and funders’ embargos; Copyright policy.
- **Preservation Policy**
Retention period; Functional preservation; File preservation; Withdrawal policy; Withdrawn items; Version control; Closure policy.

The repository policy of Byterfly prepared with *OpenDOAR* are available for the users on a web page⁶.

⁶ <http://archive.digibess.eu/policy>

5 COLLECTION HOSTED

For now, Byterfly hosts nineteen providers. Each provider can have one or more collections and choose what to share. The material is now very patchy, but we consider it an asset and it is in line with the second law: “*Every reader their book*”. We serve different users and everyone has different tastes (in fact we know we have different users because of the access statistics). The second law is closely related to the third law, “*Every book its reader*”, which means that each item has one or more readers who could find that item useful.

In our repository, there is a lot of material about sociological and economic data from Piedmont, including books from the University of Turin (belonging to the Bobbio library and the library of economics and management), periodicals from the Fiat Group (FGA Automotive), the Turin Chamber of Commerce, the Automobile Museum of Turin. Moreover, our repository hosts ancient books about theology and religion (provider Ordine dei Minimi di San Francesco di Paola), ancient volumes (sec. XVI-XVIII) about how to produce energy with water, wind and muscle power (provider Ircres-CNR), and materials, books and pictures about an International Musical festival organized in Turin and Milan every year (provider MITO Settembre Musica). The last collection in order of ingesting is a set of 57 art books (provider Fondazione 1563 for Art and Culture of Turin).

Most of the documents are in Italian, but there are also documents in Latin (ancient books), English, French, Spanish and German.

6 METADATA

Metadata furnish the description of the objects of the collection.

The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description.

The fifteen element “Dublin Core” described in this standard is part of a larger set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI). The full set of vocabularies, DCMI Metadata Terms [DCMI-TERMS], also includes sets of resource classes (including the DCMI Type Vocabulary [DCMI-TYPE]), vocabulary encoding schemes, and syntax encoding schemes. The terms in DCMI vocabularies are to be used in combination with terms from other, compatible vocabularies in the context of application profiles and based on the DCMI Abstract Model [DCAM].

The metadata we have adopted are the Simple Dublin Core as adequate information for our objects; moreover, it is possible to introduce multiple values for each element and we take advantage of this function for some terms.

Table 1. Simple Dublin Core

Term Name: contributor	
URI:	http://purl.org/dc/elements/1.1/contributor
Label:	Contributor
Definition:	An entity responsible for offering contributions to the resource.
Comment:	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Term Name: coverage	
URI:	http://purl.org/dc/elements/1.1/coverage
Label:	Coverage
Definition:	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Comment:	Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
References:	[TGN] http://www.getty.edu/research/tools/vocabulary/tgn/index.html
Term Name: creator	
URI:	http://purl.org/dc/elements/1.1/creator
Label:	Creator
Definition:	An entity primarily responsible for making the resource.
Comment:	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
Term Name: date	
URI:	http://purl.org/dc/elements/1.1/date
Label:	Date
Definition:	A point or period of time associated with an event in the lifecycle of the resource.
Comment:	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].
References:	[W3CDTF] http://www.w3.org/TR/NOTE-datetime
Term Name: description	
URI:	http://purl.org/dc/elements/1.1/description
Label:	Description
Definition:	An account of the resource.
Comment:	Description may include – but is not limited to – an abstract, a table of contents, a graphical representation, or a free-text account of the resource.

Term Name: format	
URI:	http://purl.org/dc/elements/1.1/format
Label:	Format
Definition:	The file format, physical medium, or dimensions of the resource.
Comment:	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].
References:	[MIME] http://www.iana.org/assignments/media-types/
Term Name: identifier	
URI:	http://purl.org/dc/elements/1.1/identifier
Label:	Identifier
Definition:	An unambiguous reference to the resource within a given context.
Comment:	Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
Term Name: language	
URI:	http://purl.org/dc/elements/1.1/language
Label:	Language
Definition:	A language of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as RFC 4646 [RFC4646].
References:	[RFC4646] http://www.ietf.org/rfc/rfc4646.txt
Term Name: publisher	
URI:	http://purl.org/dc/elements/1.1/publisher
Label:	Publisher
Definition:	An entity responsible for making the resource available.
Comment:	Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Term Name: relation	
URI:	http://purl.org/dc/elements/1.1/relation
Label:	Relation
Definition:	A related resource.
Comment:	Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Term Name: rights	
URI:	http://purl.org/dc/elements/1.1/rights
Label:	Rights
Definition:	Information about rights held in and over the resource.
Comment:	Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.

Term Name: source	
URI:	http://purl.org/dc/elements/1.1/source
Label:	Source
Definition:	A related resource from which the described resource is derived.
Comment:	The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Term Name: subject	
URI:	http://purl.org/dc/elements/1.1/subject
Label:	Subject
Definition:	The topic of the resource.
Comment:	Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary.
Term Name: title	
URI:	http://purl.org/dc/elements/1.1/title
Label:	Title
Definition:	A name given to the resource.
Comment:	Typically, a Title will be a name by which the resource is formally known.
Term Name: type	
URI:	http://purl.org/dc/elements/1.1/type
Label:	Type
Definition:	The nature or genre of the resource.
Comment:	Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.
References:	[DCMITYPE] http://dublincore.org/documents/dcmi-type-vocabulary/

Source: (<http://dublincore.org/documents/dces/>).

We have already tested and adopted the integration of Dublin Core with Darwin Core Metadata for specific objects concerning biology. The Darwin Core includes a glossary of terms to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. The Darwin Core is primarily based on taxa, which is their occurrence in nature as documented by observations, specimens, samples, and related information.

7 DIGITIZING, RULES AND SUGGESTIONS

Usually, the objects to be digitized are books and images; for book, we mean any kind of paper publication (issue, magazine, volume, grey literature, working paper, ...) and by image any type of photo, drawing, map, slide or similar. The born digital publications are a different case: they can be available in various formats; one of the most common is definitely the pdf.

The format chosen for file preservation is TIFF, considering its characteristics of preservation of information. This applies to both books and images as well as to digital objects, (even if the source is in PDF), that will be converted from their original format to TIFF.

The recommended resolution for digitizing images is 400-600dpi, depending on the size, while for book pages 300dpi is the standard used, adequate to obtain an excellent reading reso-

lution. If the books contain some significant illustrations, it may be advisable to perform a second separate scan at 400-600dpi and then create a collection of images that can be linked individually to the related pages of the book.

It is a good rule, especially in the case of books, not to trim the scans too much; rather, it is prudent to leave the edges (front, top and bottom cut) making the online reading more realistic. It is also advisable not to crop the central cut of the pages, so that the two side-by-side pages reproduce in the viewer the exact perception of the physical book.

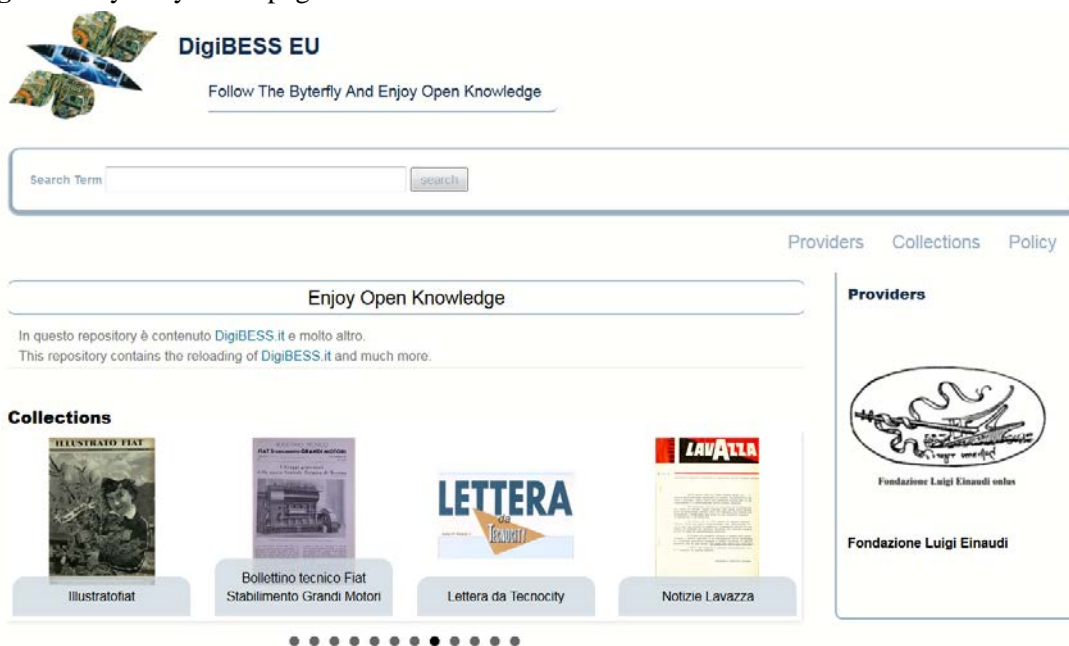
Please note that the digitization of a book must also include any blank pages of the book, as the reproduction must accurately reflect the hard copy.

The file produced by the digitization must therefore be in TIFF format, extension .tif, uncompressed and in color (RGB).

8 BROWSING AND NAVIGATE

The repository home page presents a free search textbox and a menu from which it is possible to browse the site through providers or collections. To improve the user experience, we implemented with Views two slideshow with Views in order to show collections and providers.

Figure 1. Byterfly home page



9 FINAL CONSIDERATIONS

After a year of migration work from old repositories and new ingesting, Byterfly contains more than one million pages, divided in more than 10.000 books. The objective is to keep growing in terms of contents, but also to continually improve and respect the fourth law – “*Save the time of the reader*” – by means of the creation of new ways to navigate through the repository by simplifying search function and multiplying access to the resources.

In our opinion, and following the fifth law, a repository “*is a growing organism*” and that is why we can say YES, all the five laws of library science can be applied to Byterfly.

10 REFERENCES

- Abbà S., Birello G., Vallino M., Perin A., Ghignone S., & Caciagli P. (2015). Shall we share? A repository for Open Research Data in agriculture and environmental sciences. *Bulletin EPPO Bulletin*, 45(2), 311–316. <https://doi.org/10.1111/epp.12212>
- Armeli Minicante S., Birello G., Sigovini M., Minuzzo T., Perin A. & Ceregato A. (2017). Building a Natural and Cultural Heritage Repository for the Storage and Dissemination of Knowledge: *The Algarium Veneticum* and the *Archivio di Studi Adriatici* Case Study. *Journal of Library Metadata*, 17(2), 111–125, <https://doi.org/10.1080/19386389.2017.1355165>
- Armeli Minicante S., Birello G., Ceregato A. & Perin A. (2017). Archivio Studi Adriatici (ASA) al servizio della ricerca: istruzioni per l'uso. *Rapporto Tecnico IRCrES-CNR*, n. 4/2017. <https://doi.org/10.23760/2421-5562.2017.004>
- Bertolla G., Birello G., & Perin A. (2012). Digibess, una biblioteca digitale open Source. *Biblioteche Oggi*, 30(6), 25–30.
- Birello G., & Perin A. (2017). DigiBESS reloaded and some other stories. Slides presented at Islandora Camp EU 2017, 13-15 June 2017, Delft (NL)
- Birello G., & Perin A. (2016). Gestire le risorse digitali per i beni culturali e la ricerca. Slides presented at workshop “Repository? Sì grazie!”, Area della Ricerca CNR of Turin, November, 23 2016.
- Ceregato A, Armeli Minicante S, Minuzzo T., Birello G., Perin A.(2017) *Algarium Veneticum*. Da una collezione storica alla creazione di un archivio digitale multitematico. Conference proceedings Conferenza Garr_16 Selected papers, *The CreActive Network. Uno spazio per condividere e creare nuova conoscenza*, Florence November, 30 – December, 2 2016. Isbn: 978-88-905077-6-2.